UCRL-JC--103972

DE90 012990

Projection Methods for Solving Nonlinear Systems of Equations

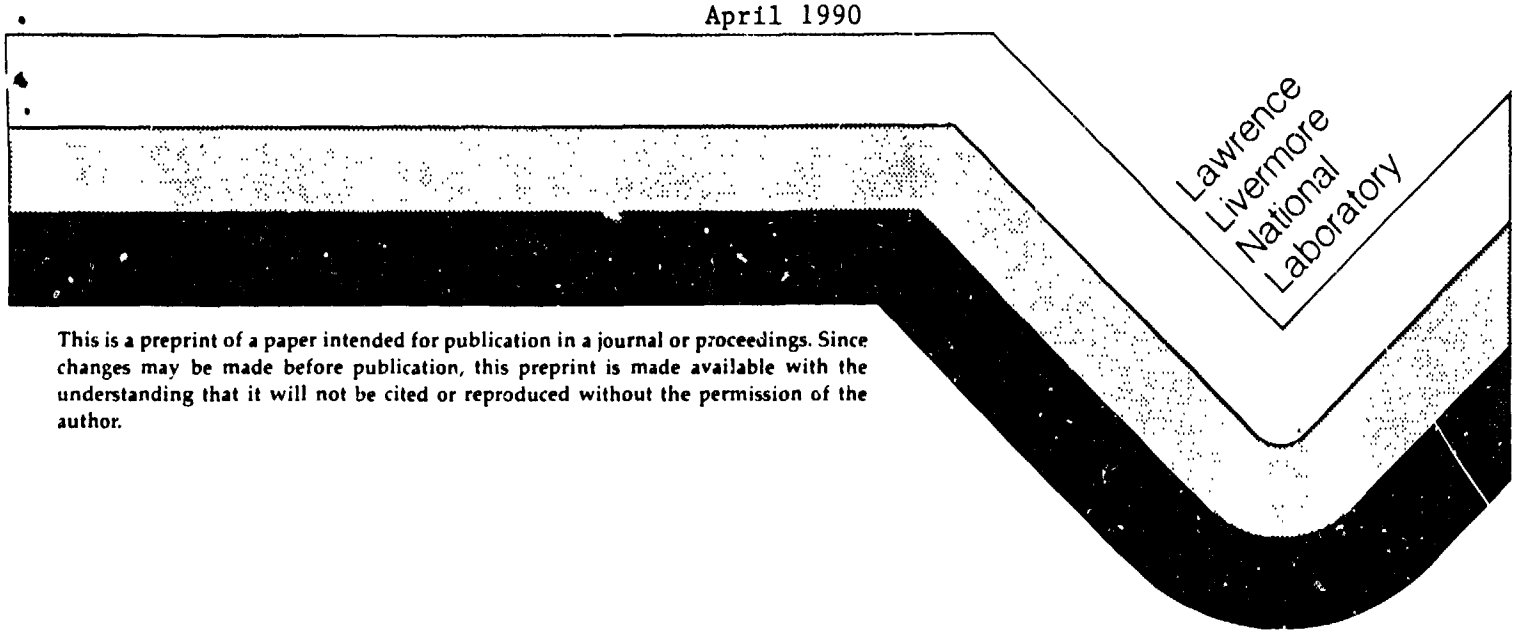
Peter N. Brown
Lawrence Livermore National Laboratory
Livermore, CA

and

Youcef Saad
Nasa Ames Research Center

This paper was prepared for the proceedings of
the NATO Advanced Research Workshop on
"Defects, Singularities, and Patterns in
Nematic Liquid Crystals,"
Orsay, France, May 28-June 1, 1990

April 1990



DISCLAIMER

Projection methods for solving nonlinear systems of equations

Peter N. Brown * Youcef Saad †

April 1990

Abstract

This paper describes several nonlinear projection methods based on Krylov subspaces and analyzes their convergence. The prototype of these methods is a technique that generalizes the conjugate direction method by minimizing the norm of the function F over some subspace. The emphasis of this paper is on nonlinear least squares problems which can also be handled by this general approach.

Keywords: Nonlinear systems; Nonlinear Projection methods; Krylov subspace methods; Inexact Newton methods; Nonlinear least squares; Conjugate gradient techniques.
AMS (MOS) Subject Classification: 65H10.

*Computing & Mathematics Research Division, L-316, Lawrence Livermore National Laboratory, Livermore, CA 94550. This work was performed under the auspices of the U.S Department of Energy by the Lawrence Livermore National Laboratory under contract W-7405-Eng-48, and supported by the DOE Office of Energy Research, Applied Mathematical Sciences Research Program.

†RIACS MS 230-5; NASA Ames Research Center, Moffett Field, CA 94035. This work was supported by the NAS Systems Division and/or DARPA via Cooperative Agreement NCC 2-387 between NASA and the University Space Research Association (USRA).

1 Introduction

The class of methods based on projections onto Krylov subspaces has proven quite effective in the solution of a wide range of problems in scientific computing. Because of the success of these methods in handling linear systems of equations and large eigenvalue problems, much effort has recently been devoted to extending their applicability to the solution of other types of problems. For example, there has been substantial progress made in using these methods for nonlinear equations arising in computational fluid dynamics [11,15]. In addition, recent work has shown how they can be used to solve equations in control theory such as Lyapunov equations [13], and there is current interest in solving time dependent partial differential equations by the method of lines [9].

The purpose of this paper is to present some of the ideas used in general nonlinear projection methods with particular attention given to nonlinear Krylov subspace methods. Algorithms of this family have been presented in [5] and a theoretical analysis was given in [4]. In this paper we extend some of the ideas in [5,4] and will emphasize a new technique suitable for solving large nonlinear least-squares problems.

When defining a nonlinear Krylov subspace method, there are two possibilities. First, one can use a globally convergent modification of Newton's iteration [8]. The linear systems that arise in the course of the Newton iteration can be solved by either a direct solver or they may be solved approximately by an iterative method. The class of methods based on the latter approach is termed *inexact Newton methods* and several such methods were considered in [1,2,3,5]. Newton's method is essentially a linearization procedure. The mapping F is locally approximated by a linear function and the resulting linear equations are solved to yield the next point. The second approach to solving nonlinear equations does not rely on linearization. Thus, fixed point iterations are inherently nonlinear as are descent methods with accurate line searches. Another well-known example is that of the nonlinear conjugate gradient iteration. In this paper we will discuss two methods, one in each of the two classes.

This distinction carries over to the definition of projection methods. In contrast with the linear case, given a subspace K there are many different ways of defining a projection process, of a Galerkin type, on the subspace. We can define a projected problem that is linear or nonlinear or a combination of both in a sense that will be clarified later. The advantages and disadvantages of each of these approaches is far from obvious. Certainly, solving a purely linear projected problem carries the advantage of simplicity. On the other hand, it may be the case that by linearizing locally, the function F will not be approximated well enough globally, and this is likely to result in a poor global convergence of the outer iteration.

We address the problem of solving a large nonlinear system, as well as that of minimizing a function from $\mathbf{R}^N$ to $\mathbf{R}^N$. The solution methods proposed can indeed handle both problems, and the theory is often identical for both cases.

2 General Nonlinear Projection Methods

In this section we introduce the basic ideas of nonlinear projection methods. We start with an overview of the different possibilities for defining nonlinear projection techniques. We will then consider the particular case where the subspaces used are Krylov subspaces.

We are interested in solving the nonlinear system

$$F(u) = 0, \tag{2.1}$$

or minimizing the function

$$f(u) = \frac{1}{2}\|F(u)\|_2^2, \tag{2.2}$$

where F is a nonlinear function from $\mathbf{R}^N$ to $\mathbf{R}^N$.

At each iteration of a nonlinear projection method we select a subspace K, and we seek an approximate solution to (2.1) or (2.2) of the form $u + \delta$, where δ belongs to the subspace K and u is the current iterate. We emphasize that the subspace K changes at every step of the nonlinear iteration. The standard case examined in [5] is when K is a Krylov subspace associated with the Jacobian of F at the current iterate. The various nonlinear projection methods we consider differ in the way the vector δ is chosen in the subspace.

For both (2.1) and (2.2), a natural choice for the next iterate is to select a vector δ in K such that

$$f(u + \delta) \equiv \frac{1}{2}\|F(u + \delta)\|_2^2 \tag{2.3}$$

is minimized. Note that although this is a nonlinear least squares problem, from a practical point of view it is much easier to solve than the original problem if the dimension m of K is much smaller than N. The motivation for this approach is that one can exploit a number of highly efficient packages, such as MINPACK, for solving least squares problems of small dimension, such as (2.3).

Let $V = [v_1, v_2, \cdots, v_m]$ be an $N \times m$ matrix whose column vectors represent an orthonormal basis of the subspace K, and write δ as $\delta = Vy$, where y is an m-vector. The function (2.3) to be minimized becomes a function of y defined by

$$g(y) = \frac{1}{2}\|F(u + Vy)\|_2^2. \tag{2.4}$$

The gradient of this function at y is given by

$$\nabla g(y) = V^T J(u + Vy)^T F(u + Vy), \tag{2.5}$$

where $J(x)$ is the Jacobian of F at the point $x \in \mathbf{R}^N$. Notice that the gradient of f is $\nabla f(u) = J(u)^T F(u)$, and so we have the simple relation $\nabla g(y) = V^T \nabla f(u + Vy)$.

A necessary, but not always sufficient, condition for y^* to be a minimum of (2.4) is that the gradient of g at y^* vanishes, i.e., we must have

$$V^T J(u + Vy^*)^T F(u + Vy^*) = 0. \tag{2.6}$$

This suggests simply solving the equations,

$$(J(u+Vy)V)^T F(u+Vy) = 0 \tag{2.7}$$

as a means for finding a minimizer of (2.4), although we know that the set of solutions of (2.7) is larger than the set of minimizers of (2.4). We refer to the above system of nonlinear equations as the set of *normal equations* for minimizing (2.4).

When solving the above normal equations, the Jacobian must be reevaluated at each new iterate and this may be uneconomical. An alternative is to freeze $J(u+Vy)V$ to be the system of vectors computed at, say, $y = 0$ and solve the set of modified equations:

$$(J(u)V)^T F(u+Vy) = 0 \tag{2.8}$$

This is a particular case of the *Petrov-Galerkin condition*

$$W^T F(u+Vy) = 0, \tag{2.9}$$

where W is an $N \times m$ matrix. Two particular cases are noteworthy:

1. $W = V$ which corresponds to the Galerkin case.
2. $W = JV$ which was naturally derived above;

When F is linear of the form $F(x) = Ax - b$, the first case corresponds to the conjugate gradient method, if F is symmetric, and Arnoldi's method when A is nonsymmetric, while the second method corresponds to the class of methods based on minimizing the residual norm, a few representatives of which are ORTHOMIN, GCR, and GMRES. See [14] for details.

Finally, one may linearize $F(u+Vy)$ in (2.9) around u and derive fully linearized techniques which correspond to solving the linear system

$$W^T[F(u) + J(u)Vy] = 0, \tag{2.10}$$

where $J(u)$ is the Jacobian of F at the current iterate u. The above linear system is m-dimensional, and will admit a unique solution if the $m \times m$ matrix $W^T J(u)V$ is nonsingular. In the particular case where $W = JV$ this condition is satisfied when the columns of JV are linearly independent. We observe that (2.10) is a way of approximately solving the Newton system $F(u) + J(u)\delta = 0$, at every step of Newton's method for solving (2.1). Thus, the fully linearized techniques are a particular case of a class of methods that are commonly referred to as *inexact Newton methods*, and have been studied in the literature, (see, e.g., [7,1,5,4]). If at every step the projected system (2.10) solves the linear system $J(u)\delta = -F(u)$ exactly, then the method becomes the standard Newton iteration. The interesting cases are again when $W = V$ and $W = JV$.

3 Fully linearized techniques

In this section we only consider the fully linearized methods in the sense defined above and summarize the results obtained in [5,4] for this case.

We start by recalling the nonlinear version of the Arnoldi (GMRES) algorithm. At every outer iteration the algorithm generates an orthonormal system of vectors v_i $(i = 1, 2, \cdots, m)$ of the Krylov subspace K^m and then builds the vector $\delta^{(m)}$. The Krylov subspace K^m is defined by

$$K^m = K(J, r, m) = \text{span}\{r, Jr, J^2r, \cdots, J^{m-1}r\},$$

for an arbitrary vector r and $N \times N$ matrix J.

Algorithm : Newton-Arnoldi (Newton-GMRES)

1. *Start:* Choose u_0 and compute $F(u_0)$. Set $n = 0$. Choose a tolerance ϵ_0.

2. *Arnoldi process:*

   - For an initial guess $\delta^{(0)}$, form $r^{(0)} = -F - J\delta^{(0)}$, where $F = F(u_n)$ and $J = J(u_n)$.
   - Compute $\beta = \|r^{(0)}\|_2$ and $v_1 = r^{(0)}/\beta$.
   - For $j = 1, 2, \cdots$, do:

     (a) Form Jv_j and orthogonalize it against the previous $v_1, \cdots, v_j$ via

     $$\begin{aligned} &h_{i,j} = (Jv_j, v_i), \quad i = 1, 2, \cdots, j, \\ &\hat{v}_{j+1} = Jv_j - \sum_{i=1}^{j} h_{i,j} v_i \\ &h_{j+1,j} = \|\hat{v}_{j+1}\|_2, \quad \text{and} \\ &v_{j+1} = \hat{v}_{j+1}/h_{j+1,j}. \end{aligned} \tag{3.11}$$

     (b) Compute the residual norm $\rho_j = \|F + J\delta^{(j)}\|_2$, of the solution $\delta^{(j)}$ that would be obtained if we stopped at this step.

     (c) If $\rho_j \leq \epsilon_n$ set $m = j$ and go to (3).

3. *Form the approximate solution:*

   Arnoldi: Define H_m to be the $m \times m$ (Hessenberg) matrix whose (possibly) nonzero entries are the coefficients h_{ij}, $1 \leq i \leq j$, $1 \leq j \leq m$ and define $V_m \equiv [v_1, v_2, \cdots, v_m]$

   - Find the vector y_m which solves the linear system $H_m y = \beta e_1$, where $e_1 = [1, 0, \cdots, 0]^T$.
   - Compute $\delta^{(m)} = \delta^{(0)} + z^{(m)}$, where $z^{(m)} = V_m y_m$, and $u_{n+1} = u_n + \delta^{(m)}$.

   GMRES: Define H_m to be the $(m+1) \times m$ (Hessenberg) matrix whose nonzero entries are the coefficients h_{ij}, $1 \leq i \leq j+1$, $1 \leq j \leq m$ and define $V_m \equiv [v_1, v_2, \cdots, v_m]$.

- Find the vector y_m which minimizes $\|\beta e_1 - H_m y\|_2$, where $e_1 = [1, 0, \ldots, 0]^T$, over all vectors y in $\mathbf{R}^m$.
- Compute $\delta^{(m)} = \delta^{(0)} + z^{(m)}$ where $z^{(m)} = V_m y_m$, and $u_{n+1} = u_n + \delta^{(m)}$.

4. *Stopping test:* If u_{n+1} is determined to be a good enough approximation to a root of (2.1), then stop, else set $u_n \leftarrow u_{n+1}$, $n \leftarrow n+1$, choose a new tolerance ϵ_n, and go to (2).

Therefore, in both Arnoldi and GMRES the outer iteration is of the form $u_{n+1} = u_n + \delta^{(m)}$ where $\delta^{(m)} = \delta^{(0)} + z^{(m)}$, with

$$z^{(m)} = V_m y_m,$$

and y_m is either the solution of an $m \times m$ linear system, for Arnoldi, or the solution of an $(m+1) \times m$ least squares problem for GMRES.

To guarantee global convergence, the usual inexact Newton methods must be modified in several possible ways. A few such modifications have been proposed in [5] and analyzed in [4]. The simplest of these involves a backtracking line-search procedure which we now consider. Given an iterate u_n we define the next iterate in the form $u_n + \lambda p_n$, where p_n is any descent direction and λ is selected by a procedure which ensures that the function f decreases sufficiently at each iteration and that the iterate makes sufficient progress towards the solution. One such procedure based on line-search backtracking is described below. The search direction p_n is provided by an approximate solution to the Newton system $J(u_n)p = -F(u_n)$, e.g., via Arnoldi or GMRES as indicated above. It is easy to show that p_n is a descent direction at u_n whenever we have

$$\|F(u_n) + J(u_n)p_n\|_2 < \|F(u_n)\|_2,$$

which means that the residual norm for the Newton system $J(u_n)p = -F(u_n)$ must be strictly reduced from that associated with $p = 0$. In particular, it is common to require that a condition of the form

$$\|F(u_n) + J(u_n)p_n\|_2 \leq \eta_n \|F(u_n)\|_2,$$

where $\eta_n \leq \eta < 1$, in the context of iterative methods.

In the procedure described below the two parameters $\theta_{\min}, \theta_{\max}$ are such that $0 < \theta_{\min} \leq \theta_{\max} < 1$, the simplest choice being $\theta_{\min} = \theta_{\max} = 1/2$. The procedure requires another parameter $\epsilon^* > 0$ which is used to essentially rescale the initial step to prevent it from from being too small.

Algorithm 3.1: General Backtracking Procedure

1. Set $\lambda = \max\{1, \epsilon^* \frac{|\nabla f(u_n)^T p_n|}{\|p_n\|_2^2}\}$.

2. If $f(u_n + \lambda p_n) \leq f(u_n) + \alpha\lambda \nabla f(u_n)^T p_n$, then set $\lambda_n = \lambda$, and exit. Else:

3. Choose $\hat{\lambda} \in [\theta_{\min}\lambda, \theta_{\max}\lambda]$; set $\lambda \leftarrow \hat{\lambda}$. Go to (2).

The following theorem [4] is a general convergence result for sequences generated by the above algorithm.

Theorem 3.1 *Let $f \equiv \frac{1}{2}\|F\|_2^2$ be differentiable and assume that its gradient is such that*

$$\|\nabla f(x) - \nabla f(y)\|_2 \leq \gamma \|x - y\|_2, \text{ for all } x, y \in \mathbf{R}^N. \tag{3.12}$$

Let p_n be such that $\|F_n + J_n p_n\|_2 \leq \eta \|F_n\|_2$ for all n, with $\eta < 1$. Further, let each iterate be chosen by the General Backtracking Algorithm. Then, either

$$\lim_{n \to \infty} f(u_n) = 0 \tag{3.13}$$

or

$$\lim_{n \to \infty} \|p_n\|_2 = \infty. \tag{3.14}$$

Moreover, superlinear convergence will essentially take place with the additional condition that η in the theorem is replaced by a sequence $\eta_n \to 0$. Global convergence of a method using a model trust region approach has also been examined in [4].

One of the most successful ways of using nonlinear Krylov subspace methods is for solving nonlinear equations in which the Jacobian of F is not available or is too expensive to compute[1]. The reason why we can still use the methods outlined above is that Krylov subspace methods do not require the Jacobian matrix J explicitly, but only its action on an arbitrary vector v. This action can be well approximated by a difference quotient of the form

$$J(u)v \approx \frac{F(u + \sigma v) - F(u)}{\sigma},$$

where u is an approximation to a solution of (2.1), and σ is some small scalar. The above observation has been exploited in several papers [15,11,10,6] to accelerate fixed-point iterations of the form

$$u_{n+1} = M(u_n)$$

by applying the above techniques to the system $F(u) \equiv u - M(u) = 0$. Typically, the Jacobian of the mapping F is a dense matrix and it may be impractical to compute it for large problems. Brown [1] has given a local convergence analysis of Newton-Krylov methods employing the above approximation.

[1]Note that the cost of producing the Jacobian may well take into account the initial programming effort.

4 Least squares projection methods

A nonlinear Krylov subspace algorithm using a least squares approach can be briefly described as follows.

Algorithm : Least Squares Krylov Subspace Method

1. *Start:* Choose the initial approximation u and compute $F(u)$.
2. *Arnoldi process:*

   Generate the orthogonal basis V_m of the Krylov subspace $K_m(J(u), v_1)$, starting with $v_1 = F(u)/\|F(u)\|_2$.
3. *Solve Projected Nonlinear Least Squares Problem:* Solve for y:

$$\min_y f(u + V_m y) \quad (\equiv \frac{1}{2}\|F(u + V_m y)\|_2^2), \tag{4.15}$$

   and set $u \leftarrow u + V_m y$.
4. Test: If satisfied then stop. Else goto 2.

There might be an ambiguity in defining the solution y in step (3) when the minimum is not unique. As will be seen later, ideally, we would like to choose the solution that is closest to u, but this may be difficult to achieve in practice. We will come back to this shortly. The attraction of the above algorithm is that there are no serious difficulties defining the iterates. In theory, the potential problems are that the solution at every step is not unique and that the sequence does not converge. However, the main disadvantage is that at every step an exact search in a whole subspace must be carried out. In practice, the least squares problems need to be solved only approximately. We require that the function f decreases, but like any other descent method, the monotonic decrease of $f(u_n)$ does not guarantee convergence of the sequence $\{u_n\}$. One way in which this may be achieved is to demand that the decrease in f at every step is sufficiently large. We would like to show a number of conditions to ensure that

$$\epsilon_n \equiv \nabla f(u_n)^T \frac{\delta_n}{\|\delta_n\|_2} \tag{4.16}$$

converges to zero, where $\delta_n = u_{n+1} - u_n$. Typically, when $\lim_{n\to\infty} \epsilon_n = 0$, the sequence u_n will converge to a solution u_* under fairly mild conditions.

If we were to perform an exact search in the subspace K then the following condition must be satisfied:

$$V^T J(u_n + V y^*)^T F(u_n + V y^*) = 0, \tag{4.17}$$

which reads

$$V^T \nabla f(u_{n+1}) = 0. \tag{4.18}$$

In other words the new gradient of f must be orthogonal to the previous subspace of projection. In practice, one may first select some u_{n+1} and then verify whether the simpler condition

$$\delta_n^T \nabla f(u_{n+1}) = 0 \tag{4.19}$$

is satisfied or nearly satisfied.

The above condition is still too stringent, since it requires solving a nonlinear equation in one variable, and we wish to find a set of conditions that the approximate solution to (4.19) must satisfy in order to ensure convergence. The next theorem establishes such a result.

Theorem 4.1 *Let $f : \mathbf{R}^N \rightarrow \mathbf{R}$, be a continuously differentiable function on $\mathbf{R}^N$ with $f(z) \geq 0$, for all $z \in \mathbf{R}^N$, and such that there exists a constant $\gamma > 0$ for which*

$$\|\nabla f(z) - \nabla f(u)\|_2 \leq \gamma \|z - u\|_2 \tag{4.20}$$

for every $u, z \in \mathbf{R}^N$. Let $\alpha > 0$, $0 \leq \mu < 1$ be given, and assume that a sequence $u_n, n = 0, 1, \cdots$, can be constructed so that at each step $u_{n+1} = u_n + \delta_n$, where $\delta_n \neq 0$ and the following conditions hold,

$$\begin{aligned} \nabla f(u_n)^T \delta_n &\leq 0 && (4.21)\\ \nabla f(u_{n+1})^T \delta_n &\geq \mu \nabla f(u_n)^T \delta_n && (4.22)\\ f(u_{n+1}) &\leq f(u_n) + \alpha \nabla f(u_n)^T \delta_n. && (4.23) \end{aligned}$$

Then,

$$\lim_{n \rightarrow \infty} \nabla f(u_n)^T \frac{\delta_n}{\|\delta_n\|_2} = 0. \tag{4.24}$$

There are a few differences between this result and that of Theorem 6.3.3 of Dennis and Schnabel [8]. First, the parameters μ and α are basically unrelated, but unlike the theorem in [8] this result does not guarantee the existence of the sequence u_n, which will be considered separately. The proof is also different and is more related to the proof of a similar result for Altman's principle given in [12]. Note that condition (4.22) can be viewed as a modification of Altman's principle.

Proof: From (4.21) and (4.22) we have

$$\begin{aligned} 0 \leq (\mu - 1) \nabla f(u_n)^T \delta_n &\leq (\nabla f(u_{n+1}) - \nabla f(u_n))^T \delta_n \\ &\leq \|\nabla f(u_{n+1}) - \nabla f(u_n)\|_2 \cdot \|\delta_n\|_2. \end{aligned}$$

Hence, using (4.20) we obtain,

$$0 \leq (\mu - 1) \nabla f(u_n)^T \frac{\delta_n}{\|\delta_n\|_2} \leq \gamma \|\delta_n\|_2. \tag{4.25}$$

From condition (4.23) and the above inequality we get

$$f(u_n) - f(u_{n+1}) \geq -\alpha \|\delta_n\|_2 \ \nabla f(u_n)^T \frac{\delta_n}{\|\delta_n\|_2} \geq \frac{\alpha(1-\mu)}{\gamma} \left[\nabla f(u_n)^T \frac{\delta_n}{\|\delta_n\|_2} \right]^2. \tag{4.26}$$

Since the sequence $f(u_n)$ is decreasing and f is bounded from below, the sequence $f(u_n) - f(u_{n+1})$ converges to zero, and as a result of (4.26) the sequence $\nabla f(u_n)^T \delta_n / \|\delta_n\|_2$ also converges to zero. □

Often, the condition (4.22) is replaced by the so-called *β-condition*

$$f(u_{n+1}) \geq f(u_n) + \beta \nabla f(u_n)^T \delta_n. \tag{4.27}$$

In fact, the same result can be shown if we replace (4.23) by (4.27).

Theorem 4.2 *Let f be a function that satisfies the same assumptions as Theorem (4.1). Let $\alpha > 0$, $0 \leq \beta < 1$ be given and assume that a sequence $u_n, n = 0, 1, ...,$ can be constructed so that at each step $u_{n+1} = u_n + \delta_n$ and the conditions (4.21), (4.23) and (4.27) hold. Then,*

$$\lim_{n \to \infty} \nabla f(u_n)^T \frac{\delta_n}{\|\delta_n\|_2} = 0. \tag{4.28}$$

Proof: We will show that the relation (4.25) is valid with μ replaced by β. Using the mean value theorem we write

$$\begin{aligned} f(u + \delta_n) &= f(u) + \nabla f(u)^T \delta_n + [\nabla f(u + \theta \delta_n)^T \delta_n - \nabla f(u)^T \delta_n] \\ &= f(u) + \beta \nabla f(u)^T \delta_n + [(1 - \beta) \nabla f(u)^T \delta_n + (\nabla f(u + \theta \delta_n)^T \delta_n - \nabla f(u)^T \delta_n)] \\ &\equiv f(u) + \beta \nabla f(u)^T \delta_n + [(1 - \beta) \nabla f(u)^T \delta_n + \|\delta_n\|_2 \zeta] \end{aligned} \tag{4.29}$$

where we have set for convenience

$$\zeta = \frac{\nabla f(u + \theta \delta_n)^T \delta_n - \nabla f(u)^T \delta_n}{\|\delta_n\|_2}$$

Note that from the assumptions we have

$$|\zeta| = |(\nabla f(u + \theta \delta_n) - \nabla f(u))^T \frac{\delta_n}{\|\delta_n\|_2}| \leq \gamma \theta \|\delta_n\|_2 \leq \gamma \|\delta_n\|_2 \tag{4.30}$$

The relation (4.29) together with the β-condition (4.27) imply that

$$(1 - \beta) \nabla f(u_n)^T \delta_n + \|\delta_n\|_2 \zeta \geq 0$$

With the inequality (4.30) this immediately yields

$$\gamma \|\delta_n\|_2 \geq - \frac{(1 - \beta) \nabla f^T \delta_n}{\|\delta_n\|_2}. \tag{4.31}$$

□

It remains to be shown that we can always select a sequence u_n provided δ_n is a descent direction and μ and α are carefully selected.

Theorem 4.3 *Let $f : \mathbf{R}^N \to \mathbf{R}$ be continuously differentiable on $\mathbf{R}^N$ with $f(z) \geq 0$ for all $z \in \mathbf{R}^N$. Let $u, \delta \in \mathbf{R}^N$ be such that $\nabla f(u)^T \delta < 0$. Then given $0 < \alpha < \mu < 1$, there exist $\lambda_u > \lambda_\ell > 0$ such that $u + \lambda\delta$ satisfies (4.22) and (4.23) for any $\lambda \in (\lambda_\ell, \lambda_u)$.*

This result is well-known and the proof may be found in Dennis and Schnabel [8]. A similar result for the β condition, can also easily be established, see for example [4]. Note the proof suggests that condition (4.23), which is the usual α-condition of Armijo and Goldstein, will always be satisfied if we replaced δ by $\lambda\delta$, with a small enough λ. As is also indicated by the proof of Theorem 4.1, the purpose of (4.22) is to prevent the step size $\lambda\delta_n$ from being too small.

The above theorem resembles Altman's principle [12] of which it is a more practical version. Convergence of sequences built from Altman's principle can be proved as a corollary to the above theorem.

Corollary 4.4 *Let $f : \mathbf{R}^N \to \mathbf{R}$ be a function satisfying the assumptions of Theorem 4.1. Let $0 < \mu < 1$ and define a sequence $\{u_n\}$ by $u_{n+1} = u_n + \lambda_n p_n$, with $\nabla f(u_n)^T p_n \leq 0$, and where λ_n is the smallest positive root of the equation in λ,*

$$\nabla f(u_n + \lambda p_n)^T p_n = \mu \nabla f(u_n)^T p_n \tag{4.32}$$

Then the conditions (4.21), (4.22) and (4.23) of Theorem 4.1 are satisfied, for any $\alpha \in (0, \mu)$, with $\delta_n = \lambda_n p_n$. In addition, $f(u_n + \lambda p_n) \leq f(u_n)$ for any λ in the interval $[0, \lambda_n]$.

Proof: The conditions (4.21) and (4.22) are obviously satisfied. To prove (4.23), we use the Mean Value Theorem which tells us that there is a certain θ, between 0 and 1 such that

$$f(u_n + \lambda_n p_n) - f(u_n) = \nabla f(u_n + \theta\lambda_n p_n)^T p_n \tag{4.33}$$

Define the function of λ

$$s(\lambda) \equiv \nabla f(u_n + \lambda p_n)^T p_n - \mu \nabla f(u_n)^T p_n.$$

We have that $s(0) < 0$. Also, since λ_n is the smallest positive root of (4.32) and $\nabla f(u_n)^T p_n \leq 0$, by continuity of $s(\lambda)$, we must have

$$\nabla f(u_n + \lambda p_n)^T p_n - \mu \nabla f(u_n)^T p_n \leq 0 \quad \text{for all } \lambda \leq \lambda_n. \tag{4.34}$$

Therefore,

$$f(u_n + \lambda_n p_n) - f(u_n) = \nabla f(u_n + \theta\lambda_n p_n)^T p_n < \mu \nabla f(u_n)^T p_n \leq \alpha \nabla f(u_n)^T p_n \tag{4.35}$$

for any $\alpha \in (0, \mu)$. The second part of the corollary follows immediately by replacing λ_n by λ in (4.33) and then exploiting (4.34). □

Assume now that we are to solve the local optimization problem at each step exactly. The search over the whole subspace will be difficult and we must add a few additional constraints.

We consider two possibilities. First, we may restrict the search to be in the level set of f at u_n, i.e., to the subset

$$L(u_n) \equiv \{\delta \in K^m | f(u_n + \delta) \leq f(u_n)\} \tag{4.36}$$

This results in,

$$f(u_{n+1}) = \min\{f(u_n + \delta) | \delta \in L(u_n)\}. \tag{4.37}$$

Note that if there is a descent direction in K_n as is always assumed, the subset L_n will not be reduced to the single point $\{0\}$. We do not know whether the minimum in the above problem is reached without any additional assumptions. Here we will assume that the initial level set $L(u_0)$ is compact so that all subsequent level sets are also compact.

A more restrictive possibility is to search only among the candidates u of K_n such that the whole interval $[0, \delta]$ is included in the level set $L(u_n)$. Here the interval $[x, y]$ denotes the set of all points of the form $tx + (1-t)y$ where $t \in [0, 1]$. This results in the definition,

$$f(u_{n+1}) = \min\{f(u_n + \delta) | [0, \delta] \subset L(u_n)\} \tag{4.38}$$

This condition on δ implies that $\nabla f^T \delta \leq 0$, in the differentiable case. Again, if there is a descent direction in K^m, the set of admissible points in (4.38) is not reduced to the single point u_n.

Although we will not show that the other assumptions of Theorem 4.1 are satisfied, we will establish that its conclusion is valid.

Corollary 4.5 *Let $f : \mathbf{R}^N \to \mathbf{R}$, a function satisfying the assumptions of Theorem 4.1 and such that the initial level set $L(u_0)$ is compact and there is a descent direction in K^m. Let $\{u_n\}$ be defined by $u_{n+1} = u_n + \delta_n$ where δ_n is defined through either of (4.37) or (4.38). Then,*

$$\lim_{n \to \infty} \nabla f(u_n)^T \frac{\delta_n}{\|\delta_n\|_2} = 0. \tag{4.39}$$

Proof: We will use an argument borrowed from [12], referred to as the "comparison principle." For this we select an arbitrary $\mu < 1$, for example $\mu = 1/2$, and an arbitrary $\alpha < \mu$, for example $\alpha = 1/4$. From u_n we create an auxiliary iterate $\bar{u}_{n+1}$ that satisfies the assumption of Corollary 4.4 with $p_n \equiv \delta_n = u_{n+1} - u_n$. Since there is a descent direction in K^m then p_n cannot be zero. We can assume, without loss of generality that $\nabla f(u_n)^T p_n \leq 0$. If not we only need to change the sign of p_n. From Corollary 4.4, the assumptions of Theorem 4.1 are satisfied and therefore the inequality (4.26) in the proof of the theorem holds, with u_{n+1} replaced by $\bar{u}_{n+1}$ (and p_n replaced by $\lambda_n p_n$ but this has no effect):

$$f(u_n) - f(\bar{u}_{n+1}) \geq \frac{\alpha(1-\mu)}{\gamma} \left[\nabla f(u_n)^T \frac{\delta_n}{\|\delta_n\|_2}\right]^2. \tag{4.40}$$

Note that by the second part of Corollary 3.3, $\bar{u}_{n+1}$ is admissible for either (4.37) or (4.38). As a result, from the definition of u_{n+1} it is clear that $f(u_{n+1}) \geq f(\bar{u}_{n+1})$, and

substituting this in (4.40) yields

$$f(u_n) - f(u_{n+1}) \geq \frac{\alpha(1-\mu)}{\gamma} \left[\nabla f(u_n)^T \frac{\delta_n}{\|\delta_n\|_2} \right]^2 . \tag{4.41}$$

Now the proof of Theorem 4.1 can be completed in the same way to establish the desired result. □

5 Numerical Experiments

As a simple example, we consider the nonlinear partial differential equation

$$-\Delta u + \alpha u_x + \lambda e^u = f \tag{5.42}$$

over the unit square of $\mathbf{R}^2$ with Dirichlet boundary conditions. This is a standard problem a simplified form of which is known as the Bratu problem [?]. After discretization by 5-point finite differencing, we obtain a large system of nonlinear equations of size N, where $N = {n_x}^2$ and n_x is the number of mesh points in each direction. The right hand side f is chosen to be the zero vector. It is known that for $\lambda \geq 0$ there is always a unique solution to the problem, see [?]. In this test we took $\alpha = 0.1$ and $n_x = 16$ yielding a nonlinear system of $N = 225$ unknowns. We tested our preliminary version of gmrls with three values of λ, namely $\lambda = +5.0, -7$, and -10.0. In the first two cases we found a solution to $F(u) = 0$ but a solution does not seem to exist for the case $\lambda = -10$. Thus, for $\lambda = -10$, our code computes the minimum of $\|F(u)\|_2$. The code incorporates an automatic switch to a nonlinear least squares projection technique, based on a simple test on the α-condition. If the GMRES solution is not admissible, a nonlinear least squares solution method, namely the routine lmdif from MINPACK, is called to minimize $f(u + V_m y)$ as was described earlier. The tolerance α for the admissibility test is set to $\alpha = 10^{-3}$.

We show the following information for each case.

Iflag	–	The termination flag (see below);
Icount	–	The total number of function calls performed;
Nfls	–	The total number of function calls that have been made by the the nonlinear least-squares routine in the projection process;
Nli	–	Total number of outer iterations;
Nlsi	–	The total number of calls to the nonlinear least-squares solver.

The stopping test involves three different criteria. The first is on the value of f. The program is stopped as soon as $f < tol1$, and *Iflag* takes the output value one. The second test relies on the value of $\nabla f(u)^T \delta / \|\delta\|_2$. The corresponmding flag is *Iflag*=2. Finally, the third criterion is on the norm of the step $\delta = u_{n+1} - u_n$ and the corresponding flag is *Iflag*=3. For all three tests the tolerance was set to 10^{-6}.

λ	Dimens.	*Iflag*	*Icount*	*Nfls*	*Nli*	*Nlsi*
-10.0	m = 10	2	3560	2949	51	49
	m = 15	2	1288	1050	14	13
-7.0	m = 10	2	629	432	17	9
	m = 15	2	436	319	7	4
5.0	m = 10	1	67	0	7	0
	m = 15	1	62	0	5	0

Table 5.1: Numerical results for the Bratu problem with different values of λ

Notice that for the harder case when $\lambda = -10$, most of the projection steps are nonlinear. The least squares problems arising in these methods are solved by MINPACK and their solution is sometimes rather expensive. Another consideration here is that we do not precondition the equations. Preconditioning, e.g., by the Laplacean could reduce the computational work in a substantial manner.

References

[1] P. N. Brown. A local convergence theory for combined inexact-Newton/ finite difference projection methods. *SIAM J. Num. Anal.*, 24:407–434, 1987.

[2] P. N. Brown and A. C. Hindmarsh. *Matrix-free methods for stiff systems of ODE'S.* Technical Report UCRL-90770, Lawrence Livermore Nat. Lab., 1984.

[3] P. N. Brown and A. C. Hindmarsh. *Reduced-Storage matrix methods in Stiff ODE systems.* Technical Report UCLR-95088, Comp. and Math. Res. Div., L-316, Lawrence Livermore Lab., Livermore Ca., 1986.

[4] P. N. Brown and Y. Saad. *Globally convergent techniques in nonlinear Newton-Krylov algorithms.* Technical Report 89-57, Research Institute for Advanced Computer Science, 1989.

[5] P. N. Brown and Y. Saad. Hybrid Krylov methods for nonlinear systems of equations. *SIAM J. Sci. Stat. Comp.*, 27., 1990. To appear.

[6] T. F. Chan and K. R. Jackson. Nonlinearly preconditioned Krylov subspace methods for discrete Newton algorithms. *SIAM J. Stat. Scien. Comput.*, 7:533–542, 1984.

[7] R. S. Dembo, S. C. Eisenstat, and T. Steihaug. Inexact Newton methods. *SIAM J. Numer. Anal.*, 18(2):400-408, 1982.

[8] J.E. Dennis and R.B. Schnabel. *Numerical Methods for Unconstrained Optimization and Nonlinear Equations.* Prentice Hall, Englewood Cliffs, NJ, 1983.

[9] E. Gallopoulos and Y. Saad. On the parallel solution of parabolic equations. In R. De Groot, editor, *Proceedings of the International Conference on Supercomputing 1989, Heraklion, Crete, June 5-9, 1989*, ACM press, 1989

[10] T. Kerkhoven and Y. Saad. *Acceleration techniques for decoupling algorithms in semiconductor simulation.* Technical Report 684, University of Illinois, CSRD, Urbana, IL., 1987.

[11] M. Mallet, J. Periaux, and B. Stoufflet. Convergence acceleration of finite element methods for the solution the euler and navier stokes equations of compressible flow. In *Proceedings of the 7-th GAMM Conference on Numerical methods in Fluid Dynamics*, page , INRIA, North-Holland, 1987.

[12] J. M. Ortega and W.C. Rheinboldt. *Iterative solution of nonlinear equations in several variables.* Academic Press, New-York, 1970.

[13] Y. Saad. *Numerical solution of large Lyapunov equations.* Technical Report 89-20, RIACS, Ms 230-5, NASA Ames, Moffett Field, CA 94035, 1989. Also in these proceedings.

[14] Y. Saad and M. H. Schultz. GMRES: a generalized minimal residual algorithm for solving nonsymmetric linear systems. *SIAM J. Sci. Statist. Comput.*, 7:856–869, 1986.

[15] L.B. Wigton, D.P. Yu, and N.J. Young. GMRES acceleration of computational fluid dynamics codes. In *Proceedings of the 1985 AIAA conference, Denver 1985*, AIAA, Denver, 1985.